CASE STUDY

Lead Scoring

Created by :
Srestha Bhar
Anurag Prakash

CASE STUDY

Overview

This case study is based on an online education service provider X that sells online courses to industry professionals where professionals who are interested in the courses land on their website and browse for courses.

How the business model works ?

1. Online platform

- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.

2. Past referrals

- Existing students who had taken the courses from the company and are satisfied with it , recommend the course to their peers.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. **The typical lead conversion rate at X education is around 30%.**

Problem Statement

How to increase the lead conversion rate?

30%

although X Education gets a lot of leads, its lead conversion rate is very poor

For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

A typical lead conversion process can be represented using the following funnel:

80%

- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- Target conversion rate is around 80%.

Lead Scoring

Problem Faced

Lead conversion rate is very low – 30%

Sollution Offered

Need to focus more on most promising leads. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

The Benefits

- Sales effort will now only be made towards the most potential leads and not wasting resources on every lead.
- Conversion rate will improve.
- Customer satisfaction will increase.

Data Set

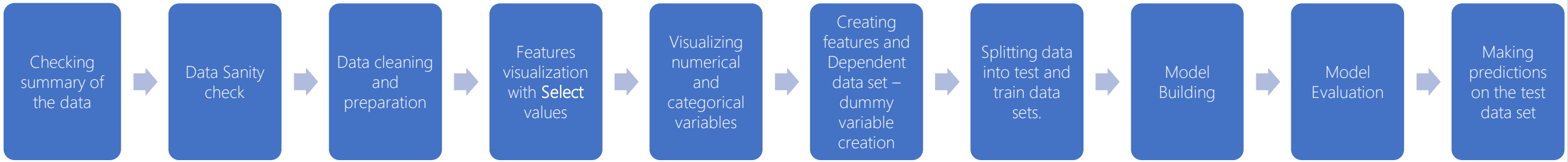
The data set contains around 9000 rows and 36 columns that has details of previous customer records.

This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Dependent Variable	Converted 0 = Not converted 1 = converted
Total features	36
Total Rows	9000
Granularity	Customer Level

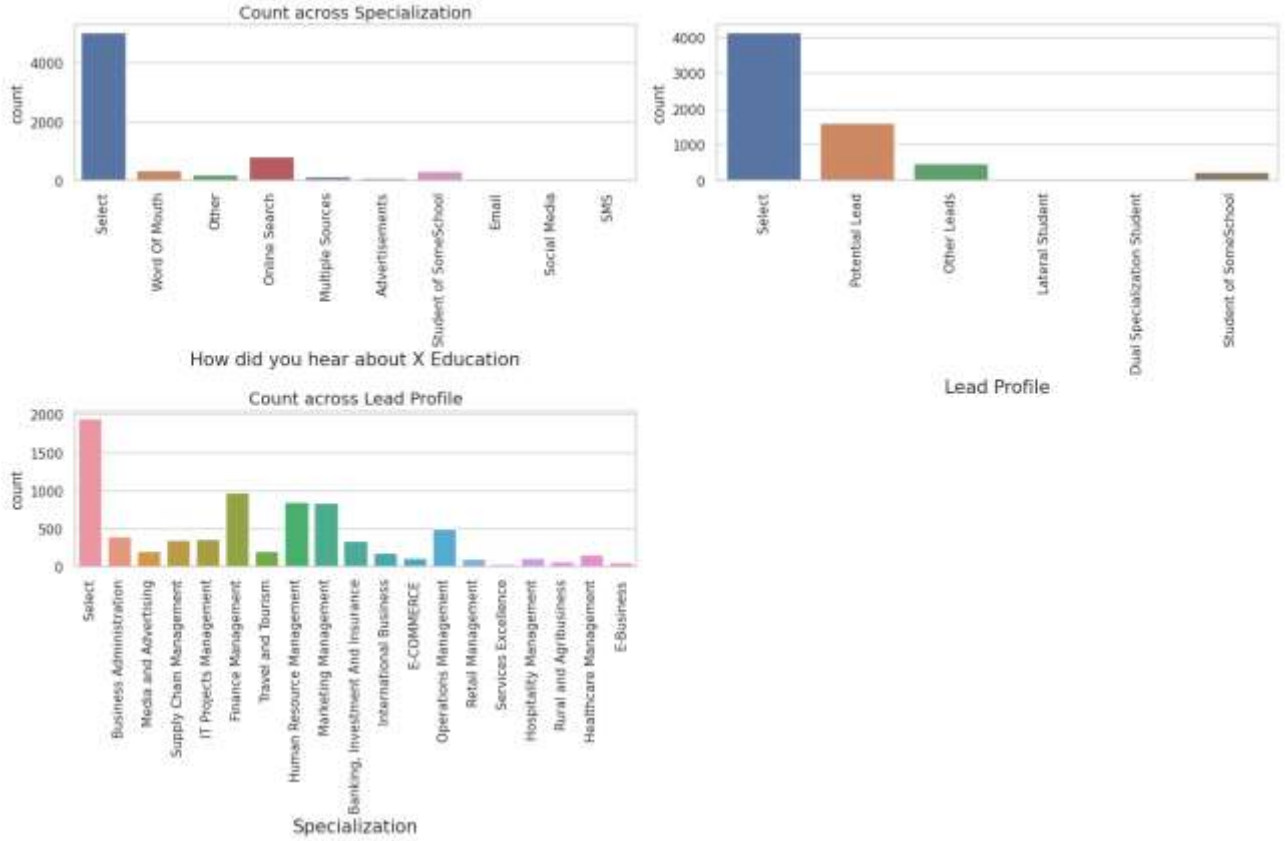
Steps for Analysis



Visualizing features with ‘Select’ values

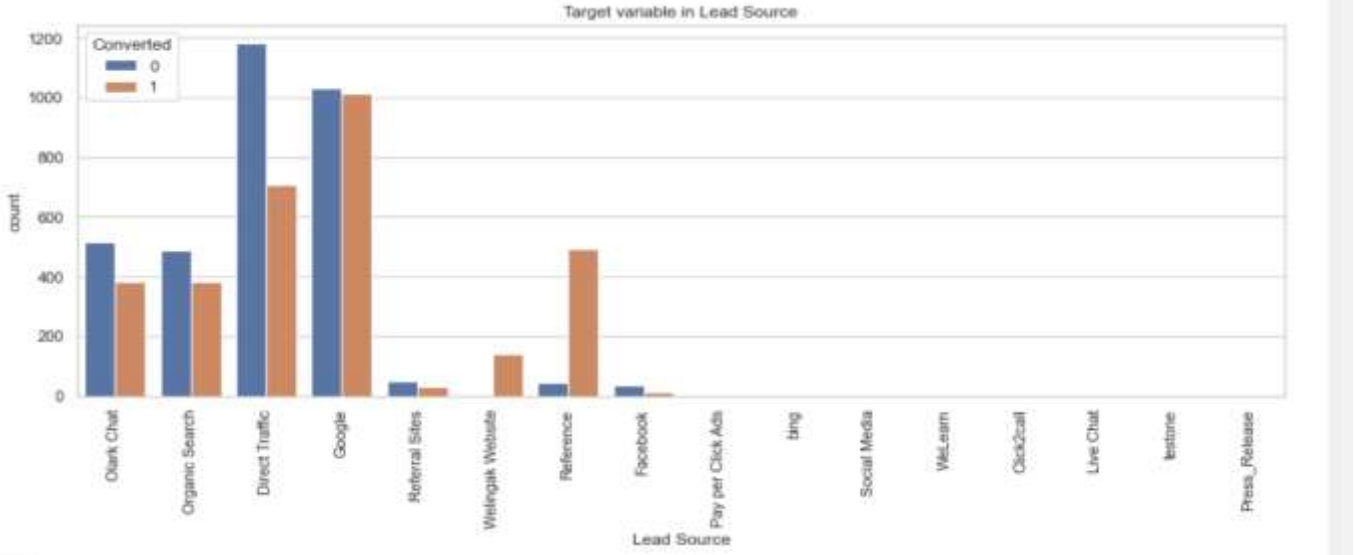
Some columns contain a level labeled 'Select,' indicating that the student did not choose a specific option for that particular column, resulting in the display of 'Select.' These instances are essentially equivalent to missing values. Therefore, it is imperative to determine the value counts of the 'Select' level in all the columns where it is present.

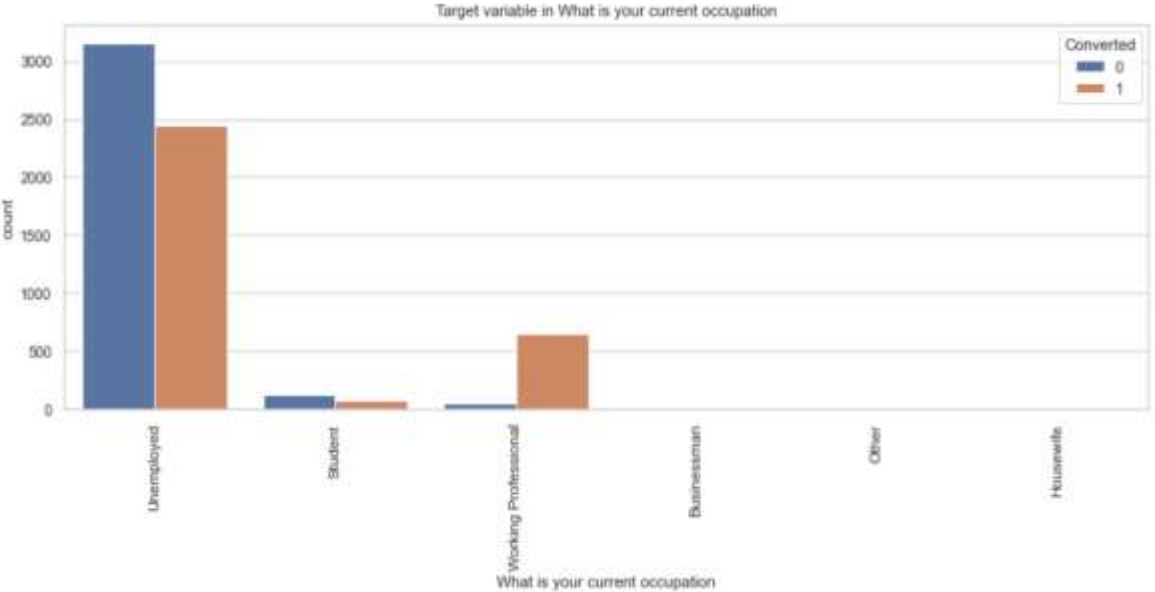
- The following three columns now have the level 'Select’ – Lead profile, How did you hear about X education, Specialization.
- Observing that the "**Lead Profile**" and "**How did you hear about X Education**" columns contain numerous rows with the value 'Select,' which **adds no value** to the analysis, it is recommended to drop these columns.

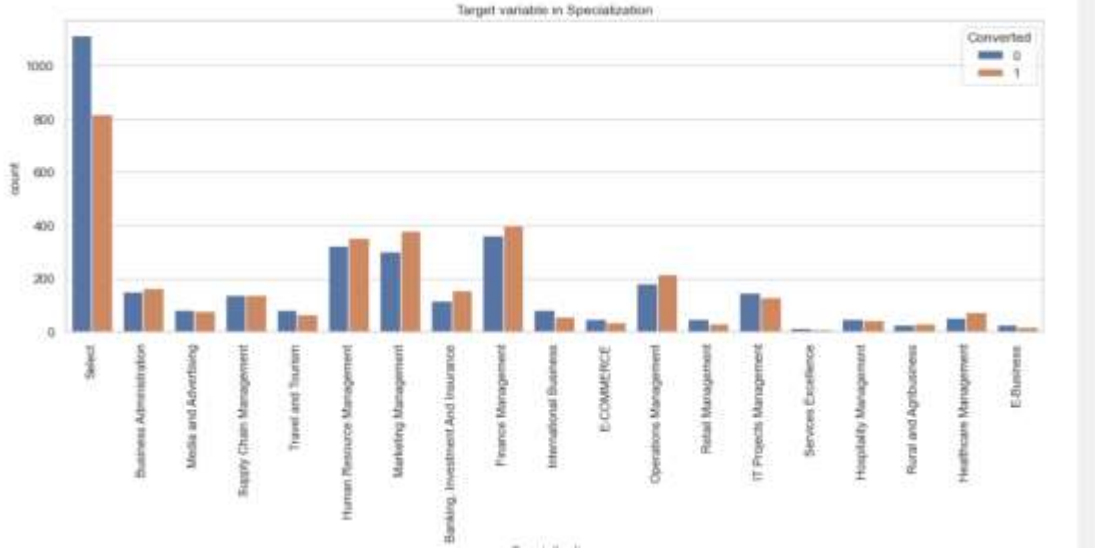


Important Business Findings

- Most of the leads that were converted got to know about the courses through **Google**.
- Most of the converted leads have experience in **finance**, **marketing** and **HR domain**.
- Majority of the leads converted were **unemployed** followed by **working professionals** and students respectively.

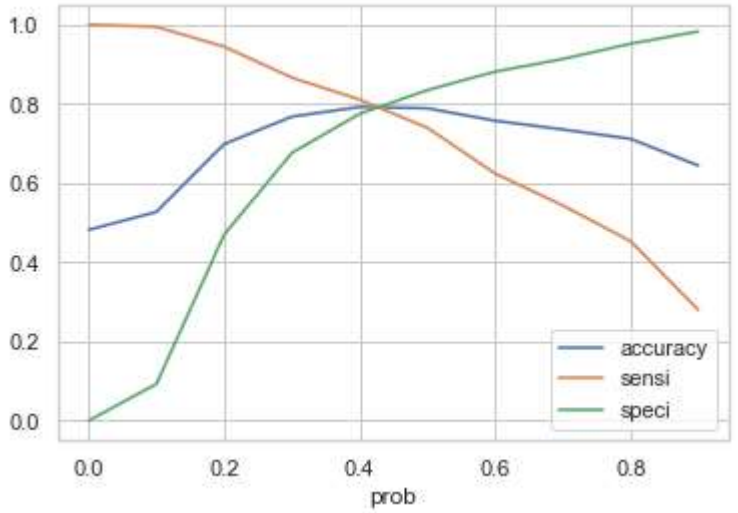






Important Business Findings

The Optimal cut-off from the ROC curve is found to be **0.42** so 0.42 is the cut off which indicates that **if conversion probability >0.42, then the lead will get converted otherwise not.**



	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted
0	0	0.300117	0	1	1	1	1	0	0	0	0	0	0	0
1	0	0.142002	0	1	1	0	0	0	0	0	0	0	0	0
2	1	0.127629	0	1	1	0	0	0	0	0	0	0	0	0
3	1	0.291558	0	1	1	1	0	0	0	0	0	0	0	0
4	1	0.954795	1	1	1	1	1	1	1	1	1	1	1	1

Important Business Findings

After implementing the final model, the features that are significant in determining the conversion status of the leads are :

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Important Business Findings

Confusion metrics- Model evaluation

- Final accuracy of the model after predicting the test data set is **~78%**
- Sensitivity/recall is ~**78%**
- Specificity is ~**79%**
- Precision is ~**80%**

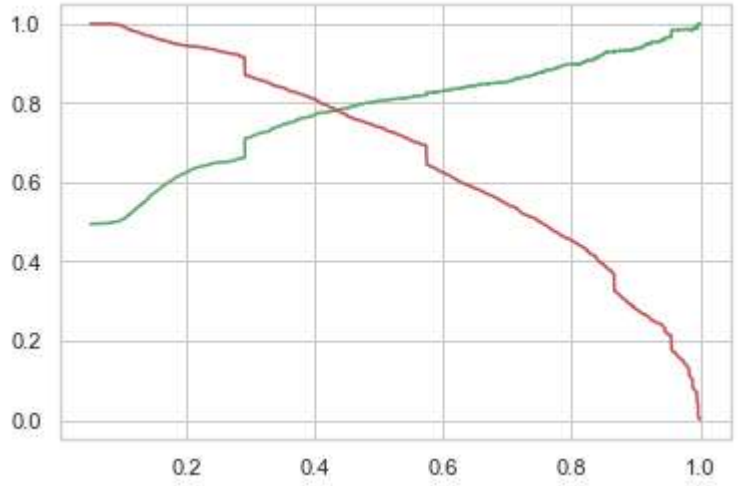
$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

Precision and recall tradeoff

In this scenario of X education business problem that is to increase the conversion rate- **Precision plays a more significant role than recall as the cost of having false positives is high**. If sales efforts are made on leads that will not purchase the service then eventually conversion rate will decrease so focus should be more on leads that will most likely get converted.



THANK YOU …